

REMARKS

Claims 1-12 were pending in the application. By this amendment, Applicants add new claims 13 and 14, which recite that the extracting of a multitude of speech features comprises using Gaussian model identities at each time frame to identify and extract features, as taught by the original Specification, for example on page 9, lines 6-12.

The Examiner has objected to the Specification due to omission of a word on page 12. The Amendment to the Specification addresses that objection.

The Examiner has rejected Claim 7, and Claims 8-12 which depend therefrom, under 35 USC 112 as indefinite. Applicants herein submit an amendment to Claim 7 to address the Examiner's concerns.

The Examiner has rejected Claims 1-12 under 35 USC 102(b) as anticipated by Vergyri. For the reasons set forth below, Applicants believe that Claims 1-14 are definite and patentable over the cited art.

The present application describes and claims a speech recognition system and method which combines features extracted from a speech training corpus, and not features that have been predetermined a priori by a human expert.

Examples of extracted features, which can number more than a million, include ranked Gaussian identities and a large speech database for direct matching. Once features have been extracted directly from the input speech, the log-linear function receives the multitude of speech features and determines a posterior probability of each of a plurality of hypothesized linguistic units given the extracted multitude of speech features. Finally, a search device analyzes the posterior probabilities determined by the log-linear function to determine a recognized output of unknown utterances.

Under the present invention, the set of features which are relevant to recognition of the input speech, and which are to be used in the log linear function, is determined based on the actual features extracted from the input speech. The set of relevant features is not predetermined, but is dynamically determined based on the acoustic input.

The number of models/features combined is not fixed but can depend on either or both of the training and the test data. For example, one may use certain acoustic phonetic features during training, but during testing, a few of these features may be unreliable due to noise -- the detectors for these features can simply abstain and our

model will do its best without these features, given the available features observed in the test data. This is possible because the maximum entropy model (a type of log-linear model, see equation 2) has a normalization factor that is dependent on the observed features. So, if, during part of a spoken utterance, there is some noise that causes some features to be unreliable, those features can be ignored and the model will make the best use of the other detected features. The parameters(λ_1) of the log-linear model in the equation 2 are obtained from the training operation. When some features used in training does not appear in testing, some parameters are not used in testing. However, there is no problem because the normalization factor ensures that equation 2 is a true probability.

In rejecting the independent claim language as anticipated, the Examiner states that:

With respect to **Claim 1**, Vergyri discloses:

A features extractor that extracts a multitude of speech features (*means for extracting multiple types of speech feature information, Pages 1, 13, 18-19, 52, and 98-99*);

A log-linear function that receives the multitude of speech features to determine a posterior probability of a hypothesized linguistic unit given the extracted multitude of speech features (*log-linear modeling function means that determines a posterior probability of a word sequence given extracted speech features from multiple information sources, Pages 100-101*); and

A search device that consults the log-linear function to determine a recognized output of unknown utterances (decoding means *performing speech recognition using the log-linear model, Pages 98, 104-109, 112; and word sequence output from a decoder, Fig. 1.1, Page 1*).

Applicants acknowledge that the Vergyri reference teaches speech recognition using features and a log linear function. However, Applicants respectfully assert that the Vergyri reference does not anticipate the invention as claimed. Vergyri does not extract features directly from the training or test speech signal. Vergyri does not dynamically determine which features are relevant based on the extracted features. Further, Vergyri teaches away from the use of a multitude of features, as detailed below.

Vergyri takes existing recognizer (or other "knowledge source" output and builds a word lattice (see: page 120). In the abstract (pp. ii-iii) of Vergyri's thesis, Vergyri equates "models" with "knowledge sources." As detailed on page 9, "[t]he new model utilizes information about the

hypothesis (or hypothesis segments) obtained from the recognizer or other available sources (such as a phone recognizer, dictionary, phone and word statistics etc.) The measurements we collect from these sources are described in Appendix B." Appendix B lists a fixed set of predetermined features which are derived from the illustrated word lattice. Vergyri maps data into a lattice and extracts the feature from the lattice. Vergyri does not extract features directly from the speech data.

Looking at Appendix B (pp. 119, et seq.), it is apparent that the measurements Vergyri is referring to are "features" that "contain information about the correctness of the hypothesized segment" (p. 18) and are not features extracted from the acoustic input. Vergyri combines a small and fixed set of scores (e.g. ac_score, lm_score, duration, num_phones, etc.) when applying a log linear function.

In contrast, the present invention combines a multitude of features that is determined by potentially a large database (of for example, Gaussian prototypes, or DTW templates, etc.) wherein the number of things to be combined is not determined a priori by the human but rather is determined based on the actual features extracted from the training data as well as the test data.

For example, as described on page 9, lines 4-10 of the present Specification, one embodiment of the invention uses ranked Gaussian identities as features. A set of Gaussian probability distributions are trained using traditional HMM technology, which for a large vocabulary system could include 300K Gaussians covering the entire acoustic space. This is very different from DMC which may combine two models (acoustic+language) or at most a few tens of predefined models which are fixed.

Another example from the Specification describes extracting features by direct matching, using dynamic time warping, to match unknown speech to a large database of reference speech segments. In this case the number of features can be extremely large to store many reference speech segments from different speakers, content, etc. resulting in potentially millions of segments. As taught on page 7 of the Specification, "[f]or example, the exemplary direct matching element may compute a dynamic time warping score against various reference speech segments in the database."

Accordingly, a big difference between Vergyri's method/system and the presently claimed method/system is in how features are extracted for use in the log linear model and on the number of lambda parameters in the log linear

model. In Vergyri's model, the set of parameters contains at most 100 (as seen in Appendix B). In the present invention, the number of parameters from the set of extracted features depends on the training data and could number in the tens of thousands or millions, or, potentially, more.

Vergyri expressly teaches that, in Equation 2.1, there are m lambda parameters, each corresponding to a "knowledge source" or model. As stated before, m is a small number under 100. The most useful measurements are given in Table 8.3 on p. 111. Other measurements are given in Appendix B. Most are derived from the word lattice (word hypothesis graph structure obtained from the search procedure of matching acoustic signal to the speech recognition models) as, for example, the score of a particular link in the lattice. Vergyri further teaches on page 101 that "[i]t is obvious that in the trivial case where the only Y_i measurements collected are the acoustic and language model log probabilities for the word link w_i and the number of words Y_0 , (8.5) becomes the usual speech recognition likelihood which uses a language model scaling factor and a word insertion penalty." As is clearly taught therein, V is not extracting features directly from the input speech, but from the lattice. Further, Vergyri is combining link

scores that are unrelated to the identity of the acoustic unit (e.g. sub-phones, phones, words). Vergyri combines link scores without considering what hypothesized unit corresponds to the link. In contrast, the present invention has different parameters lambda for different units (e.g. phones such as /a/ or /k/). In fact, the number of parameters is very large and depends on the richness of the training data (phonetics of language, variability of speakers, etc.) - the number of parameters is not pre-determined by human intuition but is rather determined algorithmically depending on the variability and size of the training data. Applicants respectfully conclude that the teachings of the Vergyri reference do not anticipate the invention as claimed.

Anticipation under 35 USC 102 is established only when a single prior art reference discloses each and every element of a claimed invention. See: In re Schreiber, 128 F. 3d 1473, 1477, 44 USPQ2d 1429, 1431 (Fed. Cir. 1997); In re Paulsen, 30 F. 3d 1475, 1478-1479, 31 USPQ2d 1671, 1673 (Fed. Cir. 1994); In re Spada, 911 F. 2d 705, 708, 15 USPQ2d 1655, 1657 (Fed. Cir. 1990) and RCA Corp. v. Applied Digital Data Sys., Inc., 730 F. 2d 1440, 1444, 221 USPQ 385, 388 (Fed. Cir. 1984). With specific reference to the language of the independent claims, Vergyri does not teach

means or steps for extracting a multitude of speech features directly from input speech; using a log linear function for determining a posterior probability of each of a plurality of hypothesized linguistic units given the extracted multitude of speech features, or determining a recognized output of unknown utterances using the posterior probabilities. Since the Vergyri reference does not teach all of the claimed features, it cannot be concluded that Vergyri anticipates the invention as claimed.

Applicants further assert that Vergyri teaches away from the invention as claimed, and cannot therefore be said to obviate the invention as claimed. On p. 15 of the thesis, Vergyri writes: “[i]t is obvious that it would be impossible to actually train one exponent λ_i for every w . Since the space I includes the acoustics, that would require at least enough acoustic data to have evidence for every word in the vocabulary (if the segments w correspond to words).” Vergyri further states on p. 100 that “[s]ince it is practically impossible to compute the normalization factor $z(\Lambda)$ in (8.1), we will treat λ_0 as we do the rest of the weights, and optimize it using training data.” Applicants respectfully note that the present invention does train one exponent λ for every w and can actually compute a normalization factor

(see: Equation 2 on page 13) to allow selective utilization of the features which are useful and relevant to the speech to be recognized.

In response to the rejections of the independent Claims 1 and 7, and the claims which depend therefrom and add further limitations thereto, as anticipated by Vergyri, therefore, Applicants request reconsideration of the rejections.

Applicants further note the following in response to the further rejections of the dependent claims.

With respect to Claim 2, Vergyri discloses:

The log linear function models the posterior probability using a log linear model (*log-linear sentence model, Page 100-101*).

Applicants reiterate that Vergyri's model is different from the invention in that there is a lambda parameter for each "knowledge source" (e.g. there are two lambda parameters if you use the two knowledge sources: (1) av_duration=expected acoustic duration of the word on the link, computed from the baseform pronunciation using the transition probabilities of the acoustic models, and (2) num_phones=number of phones in baseform pronunciation). These knowledge sources are pre-determined using human

intuition and fixed even before analyzing the training data. Hence the number of lambda parameters for the log-linear model is very small, at most equal to the number of "measurements" listed in Appendix B.

In contrast, the present invention uses a log linear model to combine a multitude of features which are determined by the training data, not human intuition. There is a lambda parameter for each feature, so the number of parameters can be hundreds of thousands, millions, or even more. Examples of features described in the patent include (1) the identities of the Gaussian models covering the acoustic space of the training data, (2) a database of speech templates used in dynamic time warping to compare with the unknown speech. Given appropriate storage and retrieval technologies, the present invention allows the use of an extremely large database to store all the available speech training data.

With respect to **Claim 3**, Vergyri further discloses:

The speech features comprise at least one of asynchronous, overlapping, and statistically non-independent speech features (*overlapping speech features, Page 18*).

On page 18, Vergyri writes:

"There can be many choices for these features, some of which may even contain overlapping information." However, on p. 110, Vergyri writes: "[t]here could be space for improving the system even more by adding more features. But this would have to be done in such a way as to avoid correlated features to interact in the system since such features would result in free parameters (extra degrees of freedom in the optimization process). A selective procedure is necessary to augment a certain set of features used with new ones that would be as uncorrelated as possible with the old features, but still contain enough information (as suggested by the correlation with the oracle measure introduced in the previous section.)"

Vergyri did not use many features (under 100) in her studies and, when contemplating using more features, explicitly stated the concern that new features must be selected to be as uncorrelated as possible with the ones already chosen. Otherwise, there is an expressed concern that the features would hurt the system by introducing too many parameters, some of which would be correlated or non-independent.

Understanding that statistical independence implies zero correlation, the present invention maintains no assumptions about statistical independence of the features,

so it is able to use a large multitude of features without worrying about the statistical independence or non-correlation of the features. Asynchronous and overlapping features may be correlated, but the invention can handle them unlike traditional Hidden Markov Models for speech recognition. Although Vergyri mentions overlapping information, her work does not anticipate the invention.

With respect to **Claim 4**, Vergyri further discloses:

At least one of the speech features extracted is derived from incomplete data (*use of a log-linear model for sparse or insufficient speech data, Page 14*).
On p. 14, Vergyri writes that “[e]rrors due to the acoustic model may be caused from bad pronunciation modeling, noisy acoustic signal, too fast or too slow speech or insufficient data for the estimation of some of the parameters.” What Vergyri refers to is a statistical fact that statistical estimation of parameters can be very poor without sufficient data. Vergyri does not, however, teach or suggest the extracting features directly from incomplete data, followed by applying the log linear function and using the resulting posterior probabilities to determine the recognized output.

The present invention is more robust to data sparseness given the large number of parameters for a maximum entropy model.

With respect to **Claim 5**, Vergyri further discloses:

The system of claim 1, further comprising a loopback (iterative processing (loopback) for error minimization (i.e., likelihood optimization), Pages 24-29 and 106-109).

What the present application teaches is that "during training, the log-linear output may be provided to the search device 3225 which can refine and provide a better linguistic unit sequence choice and a more accurate time alignment of the linguistic unit sequence to the speech. This new alignment may then be looped back to the feature extractor 3222 as FEEDBACK to repeat the process for a second time to optimize the model parameters. It should be appreciated that the initial time alignment may be bootstrapped by human annotation or by hidden Markov model technology. Thus, the model parameters corresponding to the maximum likelihood are determined as the training model parameters, and are sent to the model data element 327, where they are stored for the subsequent speech recognition operations." Vergyri does not teach or suggest loopback in

the cited passages, let alone iterative feature extraction and probabilistic probability determination.

With respect to **Claim 6**, Vergyri further discloses:

The features are extracted using direct matching between test data and training data (*extracted feature measurements resulting from speech recognition matching, Pages 9 and 109-126*).

In response to this, Applicants reiterate that Vergyri's model is different from the invention in that there is a lambda parameter for each "knowledge source". The knowledge sources are pre-determined using human intuition and fixed even before analyzing the training data. Hence the number of lambda parameters for the log-linear model is very small, at most equal to the number of "measurements" listed in Appendix B. In contrast, the present invention uses a log linear model to combine a multitude of features which are determined by the training data, not human intuition. There is a lambda parameter for each feature, so the number of parameters can be hundreds of thousands, millions, or even more. For direct matching, one exemplary embodiment would be to keep all the training data speech segments in a large database (instead of commonly just keeping estimated statistical models) and

match the speech templates to the unknown speech segments.

The foregoing is neither taught nor suggested by Vergyri.

In summary, the present patent application describes and claims a system and method to combine features that can be extracted from the speech training corpus and not necessarily determined *a priori* by a human expert. Examples of features include ranked Gaussian identities and large speech database for direct matching. The number of features (and corresponding lambda parameters for the log-linear model) can be more than a million. Vergyri's model does not extract features directly from the speech and Vergyri's method uses exactly the same number of models in testing as during training (i.e., the number of lambda parameters is used (e.g. m is a fixed number in Equation 2.2)). In contrast, the present invention can use a variable number of features depending on what was actually observed in the test speech signal. Thus the number of features used during testing need not be the same as that used in training. For a speech frame or linguistic unit, it may be only a few features (and corresponding lambda parameters) will be used out of the millions estimated during training. As another example, the claimed invention can use speech distinctive feature detectors that are not conventionally used in Hidden Markov Model based

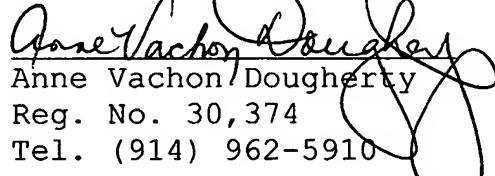
speech recognition. It is well known by speech scientists that certain features like aspiration noise can distinguish between stop closures like /b/ and /p/ in clean acoustic environments. However, in noisy environments, the aspiration noise can be masked, so a better feature is the voice-onset-time. During training using clean speech, one can use both aspiration noise and voice-onset-time as features and estimate appropriate lambda parameters. In testing, if it is noisy, the aspiration noise detector can simply abstain (not fire) and only the voice-onset-time and other appropriate features that are actually observed can be used in the log-linear model. Because the model is appropriately normalized over the observed features and contexts, a true probability model is obtained just based on what is observed, regardless of how many or how few features were activated (how many lambda parameters used). Thus the number of features (and lambda parameters) used in testing need not be the same as used in training. No such mention of this novel operating mode was mentioned in Vergyri's thesis.

Since the Vergyri reference does not teach each and every claim feature, Applicants respectfully request entry of the amendments, reconsideration of the anticipation rejections under 35 USC 102, and issuance of the claims.

Respectfully submitted,

Axelrod, et al

By:


Anne Vachon Dougherty
Reg. No. 30,374
Tel. (914) 962-5910